Assessing the Comprehensibility of Automatic Translations (ArisToCAT)

Lieve Macken, Margot Fonteyne, Arda Tezcan and Joke Daems
LT[3], Language and Translation Technology Team
Ghent University
Belgium
Lieve.Macken@ugent.be

Abstract

The ArisToCAT project aims to assess the comprehensibility of 'raw' (unedited) MT output for readers who can only rely on the MT output. In this project description, we summarize the main results of the project and present future work.

1 Introduction

Machine translation (MT) systems cannot guarantee that the text they produce will be fluent and coherent in both syntax and semantics. Errors occur frequently in machine-translated text, leaving the reader to guess parts of the intended message. With the arrival of neural machine translation (NMT), however, the quality of machine translation has increased significantly. As such, machine translation is becoming an attractive solution to deal with the increased need for translated content. This could mean that, in the near future, readers will be more often confronted with 'raw' (unedited) MT output.

2 Quality of MT output

To assess the quality improvements in MT, we compared the quality of three different MT systems for English–Dutch: a commercial neural system, a phrase-based system and a predominantly rule-based system. We used Web-Anno[1] as annotation tool and adopted a two-step approach to annotate all errors in the MT output. In a first step, only the target text was visible and we marked all fluency errors; in a second step all accuracy errors were labelled in both source and target text and were linked. Van Brussel et al. (2018) found that the neural system, in general, outperformed the phrase-based and rule-based systems when considering fluency. The output of the neural system contained fewer grammatical errors and hardly any spelling mistakes. For accuracy, the improvements of NMT are less apparent. The target sentence does not always contain traces of the errors or clues of omissions, which might have an impact on the comprehension.

3 Reading comprehension tests

In a pilot study, Macken and Ghyselen (2018) selected three texts of the English MT Evaluation version of the Corpus of Reading Comprehension Exercises (Scarton and Specia, 2016) and set up a reading comprehension test for both human translated and raw MT texts. Ninety-nine participants were asked to read the translation very carefully after which they had to answer comprehension questions without having access to the translated text. Human translations received the best overall clarity scores, but the reading comprehension tests provided much less unequivocal results.

4 Comprehensibility of newly invented words in NMT output

NMT systems occasionally generate non-existing words, i.e words that are not part of the vocabulary of the target language and were thus invented by the NMT system. In cases in which readers only have access to the MT output without the source text, such non-existing words can affect comprehension. There are several reasons why an NMT system creates new non-existing words. One reason is that, although NMT systems have

[1]https://webanno.github.io/webanno/

Martins, Moniz, Fumega, Martins, Batista, Coheur, Parra, Trancoso, Turchi, Bisazza, Moorkens, Guerberof, Nurminen, Marg, Forcada (eds.)
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, p. 485–486
Lisboa, Portugal, November 2020.

made huge progress, they sometimes still generate a too literal translation for different types of multi-word expressions such as compounds, another reason, specific for NMT systems, is that they operate at sub-word level to reduce vocabulary size. Macken et al. (2019) set up an experiment in which eighty-six participants were given 15 non-existing words (5 single words and 10 noun compounds) and were either asked to describe the meaning of these words or to select the correct meaning from a predefined list. The words were presented either in isolation or in sentence context. Non-existing words indeed impaired comprehension as on average in 60% of the cases the participants gave a wrong answer. Sentence context, however, made it easier for the participants to determine the meaning of the non-existing word as the percentage of wrong answers is much higher (77%) when the words were presented in isolation.

5 NMT for literary translation

To assess whether, with the improved quality, NMT systems are able to produce high-quality translations for more creative text types such as literature, we translated Agatha Christie's novel *The Mysterious Affair at Styles* with Google's NMT system into Dutch and applied the two-step error annotation. Fonteyne et al. (2020) found that 44% of the MT sentences did not contain any errors. The accuracy subcategory *mistranslation* was the most frequent error type encountered in the novel, followed by the fluency subcategories *coherence* and *style & register*. Tezcan et al. (2019) further investigated how the MT version differs from the published professionally human-translated (HT) Dutch version of the book. Measures of lexical richness (type-token ratio and mean segmental type-token ratio) gave inconclusive results. They also looked at word translation entropy (Carl et al., 2016), which indicates the degree of uncertainty to choose a particular translation from a set of target words based on the number and distribution of different translations that are available for a given word in a given context and found that the average word translation entropy scores were higher in HT than in MT, meaning that there was more variety in the translations in HT. At the syntactic level, the MT generally follows more closely the structure of the source sentence compared to the HT version.

6 Future work

In future work, we will set up eye-tracking experiments and expand the Ghent Eye-Tracking Corpus (Cop et al., 2017) with eye-tracking data for the NMT version of the novel. This will allow us to analyse to what extent MT impacts the reading process, and which errors impact this reading process the most.

Acknowledgements: The ArisToCAT project is a four-year research project (2017–2021) funded by the Research Foundation – Flanders (FWO) – grant number G.0064.17N assigned to prof. dr. Lieve Macken. All research is carried out at the University of Ghent (Belgium).

References

Carl, M., M. Schaeffer and S. Bangalore 2016. The CRITT Translation Process Research Database. In *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB.*, Cham, Heidelberg, New York, Dordrecht, London: Springer, 13–54.

Cop, U., N. Dirix, D. Drieghe and W. Duyck 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

Fonteyne, M., A. Tezcan, and L. Macken. Accepted. Literary MT under the Magnifying Glass: Assessing the Quality of an NMT-Translated Agatha Christie Novel. *Proceedings of LREC 2020*

Macken, L. and I. Ghyselen. 2018. Measuring comprehension and acceptability of neural machine translated texts: a pilot study. *Proceedings of the 40th Conference Translating and the Computer*, London, UK, pp. 120–126.

Macken, L., L. Van Brussel and J. Daems 2019. NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. *Computational Linguistics in the Netherlands Journal*, 9:67–80.

Scarton, C. and L. Specia. 2016. A reading comprehension corpus for machine translation evaluation. *Proceedings of LREC 2016*, Portorož (Slovenia).

Tezcan, A., J. Daems, and L. Macken. 2019. When a 'sport' is a person and other issues for NMT of novels. *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland: European Association for Machine Translation, pp. 40–49

Van Brussel, L., A. Tezcan and L. Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. *Proceedings of LREC 2018*, Miyazaki, Japan.